1.2 Description of Work

The objective of this R & D activity is the development of a scalable SMP cluster-based system, including all hardware and software, that will provide sufficient capacity and capability to develop, debug, and execute large simulation codes at the TeraFLOP/s performance level in support of SBSS.

The Subcontractor shall:

1.2.1 Detailed Project Plan (MR)

Provide a detailed project plan for the ASCI Blue development project.

1.2.2 Execute Development Plan (MR)

Accomplish the milestones of the Subcontract; accelerate the development of hardware and software for the purpose of achieving the stated Tera-scale performance goals.

1.2.3 Install ID System(s) (MR)

Configure, install, support and maintain the Initial Delivery (ID) system, including hardware and software, at LANL or LLNL. At the option of the University, install an additional ID system.

1.2.4 Technology Refresh (TR)

Provide hardware and software technology refresh to those systems as available.

1.2.5 On-site Support (MR)

Provide on-site assistance to Laboratory applications developers.

1.2.6 Scalable Development Environment Goal (MR)

Demonstrate the scalability of essential software capabilities in the application development environment across the clustered SMP system.

1.2.7 Sustained TeraFLOP Performance Goal (MR)

Demonstrate sustained TeraFLOP/s performance on the sPPM Science-Based Stockpile Stewardship application.

1.2.8 Three Peak TeraFLOP Performance Goal (MR)

Demonstrate a machine with the sum of peak and sustained performance of at least four (4.0).

1.2.9 Install Sustained Stewardship TeraFLOP (SST) System (MR)

Configure, install, support, and maintain the TeraFLOP/s scalable clusters at LANL or LLNL.

Clusters of SMP systems that can address the anticipated stockpile stewardship application requirements are anticipated to scale according to the following ratios. These ratios were developed through the Laboratories' multi-year experience with addressing challenging computational problems on high performance computing systems from a variety of application domains.

1 FLOP/s peak performance /
.5 - 1 byte memory size /
10 -100 byte disk storage /
4-16 byte per second L1-L2-cache bandwidth /
1-3 byte per second memory bandwidth /
0.1-1.0 bit per second communications bandwidth /
0.01 - 0.1 byte per second disk bandwidth

Building and delivering a Tera-scale computing resource is a daunting task. Within the context of a research and development contract it is anticipated that a well balanced hardware approach will follow the following four notional phases: 1) an Initial Delivery (ID) system for ASCI application code development; 2) technology refresh as the ID system ages and fruits of the ASCI project become available; 3) a TeraFLOP/s system with a peak plus sustained performance of at least four (4.0) TeraFLOP/s; and 4) memory upgrade.

Due to SBSS programmatic requirements, the University has developed a general schedule for delivery of hardware and software assets. Of prime consideration are the decoupling of compute and memory SST deliveries and the timeliness of an Initial Delivery (ID) system for code development.

Target Delivery Date	**System**	**Metrics**
90 days after contract award	Initial Delivery (1st System)	100 GFLOP/s Peak Performance, 50 GB Main Memory, 2.5 TB RAID Disk
90 days after contract award	Initial Delivery (Optional 2nd System)	100 GFLOP/s Peak Performance, 50 GB Main Memory, 2.5 TB RAID Disk
2Q CY 1998	SST Delivery	Sustained sPPM TFLOP/s performance. Memory to Peak FLOP/s ratio is 0.167. At least 25 TB RAID Disk
2Q CY 1998	SW Scalability	Scalability of the application development environment tools across the clustered SMP system.
Q1 CY 1999	Memory Upgrade	Increase memory to Peak FLOP/s of SST from 0.167 up to 0.5. Additional 50 TB RAID Disk.

the objectives of this project and Subcontractor's Research and Development Plan, which the Subcontractor believes will be of benefit to the project.

4.1 SST Hardware High-Level Requirements

4.1.1 Scalable Cluster of SMPs

4.1.1.1 Sustained Stewardship TeraFLOP SMP Scalable Cluster (MR)

The Subcontractor shall provide a Sustained Stewardship TeraFLOP/s (SST) system composed of multiple Shared memory Multi-Processors (SMPs) connected via a scalable intra-cluster communications technology. The system shall have a peak plus sustained performance on the sPPM benchmark of at least four (4.0) TeraFLOP/s. The Subcontractor shall provide a best effort to obtain a sustained performance of one (1.0) TeraFLOP/s ($1.0x10^{12}$ floating point operations per second) on the sPPM benchmark.

Example: If "p" is the peak performance of the system and "s" is the sustained on sPPM performance of the system, and if we define the machine efficiency as e = s/p, then the above equation becomes:

p(e) 3 4.0/(1.0+e)

Hence,

p(1/2) 3 4/1.5 2.67, p(1/3) 3 4/1.33 = 3.0 and p(1/5) 3 4/1.2= 3.3

4.1.1.2 SST Component Scaling (MR)

In order to provide the maximum flexibility to the Subcontractor in meeting the goals of the ASCI project the exact configuration of the SST SMP scalable cluster is not specified. Rather the SST configuration is given in terms of lower bounds on component attributes. The SST SMP scalable cluster configuration shall meet or exceed the following parameters:

- Memory Size 3 0.5 TB
- Disk Space 3 75 TB
- Cache Bandwidth/Peak FP (Byte/s/FLOP/s) 3 4
- Memory Bandwidth/Peak FP (Byte/s/FLOP/s) 3 1
- Intra-Cluster Network Bi-Section Bandwidth 3 0.5 Tb/s
- System Peak Disk I/O Bandwidth 3 90 GB/s

4.1.1.2.1 Additional Intra-Cluster Network Bi-Section Bandwidth (TR)

The Subcontractor may provide a configuration identical to that specified in Requirement 4.1.1.2, but with an Intra-Cluster Network Bi-Section Bandwidth 3 1.5 Tb/s.

4.2.3.3 Accounting for System (Root) Usage (TR)

Resources used by root processes that are not otherwise considered to be part of another job and by the operating system itself may be accountable. In addition, resources not used may be accountable. Accounting for these resources may be accomplished through the same interfaces as are provided for all other jobs. Accounting for operating system resource usage may be done through a pseudo-job known as the “kernel” job. Accounting for root processes’ resource usage may be done through one or more pseudo-jobs known as “system” jobs. Finally, accounting for unused resources may be done through a pseudo-job known as the “idle” job.

4.2.3.4 Cluster Wide Job Management (TR)

A job may be a cluster wide abstraction similar to a POSIX session, with certain characteristics and attributes. Commands may be available to manipulate a job as a single entity (including kill, modify, query characteristics, and query state). The characteristics and attributes required for each session type are as follows: 1) interactive session: an interactive session include all cluster wide processes executed as a child (whether direct or indirect through other processes) of a login shell and include the login shell process as well. Normally, the login shell process exists in a process chain as follows: init, inetd, [telnetd | rlogind | xterm | cron], then shell. 2) batch session: a batch session includes all cluster wide processes executed as a child (whether direct or indirect through other processes) of a shell process executed as a child process of a batch system shepherd process, and includes the batch system shepherd process as well. 3) ftp session: an ftp session includes an ftpd and all its child processes. 4) kernel session: all processes with a pid of 0. 5) idle session: this session does not necessarily actually consist of identifiable processes. It is a pseudo-session used to report the lack of use of resources. 6) system session: all processes owned by root that are not a part of any other session.

4.2.3.5 Cluster Wide Job Scheduling (MR)

The Subcontractor shall provide a capability so that a user can submit, either interactively or through the batch environment, a job that spans any subset of user accessible processors, and to schedule multi-thread, multi-process, message-passing jobs using configurable load levels. These configurable load levels shall be defined on CPU load, available memory, paging rate (if applicable), available swap and temporary disk space.

4.2.3.5.1 Cluster and SMP Gang Scheduling (TR)

The Subcontractor may provide a capability to gang schedule threads and processes from a single user job within an SMP and across SMPs. That is, when a user job is scheduled to run, the gang scheduler may contemporaneously allocate to CPUs all the threads and processes within that job (either within an SMP or within the cluster of SMPs). This scheduling capability may control all threads and processes within the SMP cluster environment.

4.2.9.1 Login Information (TR)

Users may be notified upon successful login of the following information: date and time of last successful login; and where the operating system provides the capability, number of unsuccessful attempts.

4.2.9.2 Audit Capability (MR)

A record of each user login and logoff shall be maintained. In addition, the following information shall be maintained as an audit record: use of authentication changing procedures; unsuccessful logon attempts; and blocking of a user ID and the reason for the blocking.

4.2.10 Compliance with DOE Security Mandates (MR)

DOE Security Orders have changed over time. From time to time these mandates cause LANL and LLNL to fix bugs or implement security features in vendor operating systems and utilities. We require, for computer security purposes only on this contract, that the Subcontractor either provide to the University the operating system and utilities source code on a demand basis or make needed enhancements or bug fixes as required.

4.2.11 Cluster Environment On-Line Document (TR)

The Subcontractor may supply hardcopy and on-line documentation for all major hardware and software subsystems. The on-line documentation may be readable by all users utilizing an X11-based graphical user interface and standard terminal interfaces.

4.2.12 SST Applications Development Support (MR)

The Subcontractor shall supply at least two on-site analysts to provide expertise to the University code development teams in the areas of software development tools, parallel applications libraries and applications performance. The proposed system will be installed in a classified area at the Laboratory and so analyst personnel shall obtain DOE Q clearances.

End of Section 4

5.2.8.4 Graphical User Interface API (MR)

The Subcontractor shall provide the standard X11R5 and Motif 1.2, or current versions, applications, servers and API libraries.

5.2.8.5 Math Libraries (TR)

The Subcontractor may supply highly optimized mathematical libraries, callable from all baseline languages, for serial (single processor) SMP and cluster wide applications.

5.2.9 SST Software Environment Features for ID (TR)

The ID system may present a similar general software environment as the SST does. The University is aware that the specifics (e.g., clustering software, intra-SMP code development tools, software scalability) between the two systems may differ. Indicate which of the SST Software target requirements in section 4.2 may be delivered with the ID (or shortly thereafter as part of a software technology refresh program).

5.2.10 Compliance with DOE Security Mandates (MR)

DOE Security Orders have changed over time. From time to time these mandates cause LANL and LLNL to fix bugs or implement security features in vendor operating systems and utilities. We require, for computer security purposes only on this contract, that the Subcontractor either provide to the University the operating system and utilities source code on a demand basis or make needed enhancements or bug fixes as required.

5.2.11 Cluster Environment On-Line Documentation (TR)

The Subcontractor may supply hardcopy and on-line documentation for all major hardware and software subsystems. The on-line documentation may be readable by all users utilizing an X11-based graphical user interface and standard terminal interfaces.

5.2.12 ID Applications Development Support (MR)

The Subcontractor shall supply at least one on-site analyst to provide expertise to the University code development teams in the areas of software development tools, parallel applications libraries and applications performance. The proposed system will be installed in a classified area at the Laboratory and so analyst personnel shall obtain DOE Q clearances.

5.3 Performance of the ID System

The seven benchmark programs described below may be executed by the Subcontractor for the purpose of measuring the execution performance and compiler capabilities of the ID system. Each of the benchmark programs represents a particular subset and/or characteristic of the expected ASCI workload, which consists of solving complex scientific problems using a variety of state-of-the-art computational techniques. The general requirements and constraints outlined below shall apply to all of the benchmark codes. Additional requirements and/or constraints found in individual benchmark readme files shall apply to that individual benchmark.

6.2 Project Milestones

Because of the need to meet Science-Based Stockpile Stewardship goals as quickly as possible, the project schedule and milestones are of critical importance. Meeting the following milestones is critical to the success of the project; earlier is much better.

The implementation will entail the installation of local clusters at Laboratory sites. Each cluster will be assembled from individual SMPs that are interconnected with a high-speed, low-latency interconnect supplied by the Subcontractor. These clusters will be connected to the site's local campus network and to the wide area network that interconnects the three Laboratories. Access to the resources will be provided locally via the site's existing campus networks and remotely through the ASCI-supplied WAN.

6.2.1 Detailed Project Plan (MR)

The Subcontractor shall provide a detailed Project Plan 30 days after contract award.

6.2.2 Initial Delivery (ID) System (MR)

The Subcontractor shall install and support an initial code development system containing 50 Gigabytes (GB) of memory, 2.5 TB of disk configured for RED/BLACK operations and capable of 100 GigaFLOP/s (GFLOP/s) of peak performance at LANL or LLNL, as directed, 90 days after contract.

6.2.3 Second Initial Delivery (ID) System (MO)

At the option of the University, the Subcontractor shall install and support a second initial code development system containing 50 Gigabytes (GB) of memory, 2.5 TB of disk configured for RED/BLACK operations and capable of 100 GigaFLOP/s (GFLOP/s) of peak performance at LANL or LLNL, as directed, 90 days after contract award.

6.2.4 ID Applications Development Support (MR)

The Subcontractor shall supply at least one on-site analyst to provide expertise to the University code development teams in the areas of software development tools, parallel applications libraries and applications performance at the time of ID system delivery. The proposed system will be installed in a classified area at the Laboratory and so analyst personnel shall obtain DOE Q clearances.

6.2.5 FY97 Plan and Review (MR)

The Subcontractor shall provide a detailed plan of activities and deliverables for fiscal year 1997 for University review and approval in the first quarter of FY97.

6.2.6 Technology Refresh (TR)

The Subcontractor may provide hardware and software technology updates between the installation of the ID system and the installation of the SST that would significantly improve the hardware and software environment. Hardware technology updates may meet or exceed the following component scaling parameters:

- Memory Size/Peak FP (Byte/FLOP/s) 3 0.5
- Disk Space/Peak FP (Byte/FLOP/s) 3 25
- Cache Bandwidth/Peak FP (Byte/s/FLOP/s) 3 4
- Memory Bandwidth/Peak FP (Byte/s/FLOP/s) 3 1
- Intra-Cluster Network Bi-Section Bandwidth/Peak FP (Bits/s/FLOP/s) 3 0.167
- System Peak Disk I/O Bandwidth/Peak FP (Byte/s/FLOP/s) 3 0.03

6.2.7 FY98 Plan and Review (MR)

The Subcontractor shall provide a detailed plan of activities and deliverables for fiscal year 1998 for University review and approval in the first quarter of FY98.

6.2.8 SST Applications Development Support (MR)

The Subcontractor shall supply at least two on-site analysts to provide expertise to the University code development teams in the areas of software development tools, parallel applications libraries and applications performance at the three months prior to the SST system delivery. The proposed system will be installed in a classified area at the Laboratory and so analyst personnel shall obtain DOE Q clearances.

6.2.9 Scalable Development Environment Demonstration (TR)

The Subcontractor may demonstrate the scalability of essential software capabilities in the application development environment across the clustered SMP system in mid CY 1998. Specifically, the Subcontractor may use the sPPM demonstration code on the full cluster to demonstrate debugger, event tracing and performance statistics capabilities.

6.2.10 Sustained Stewardship TeraFLOP (SST) Demonstration (TR)

The Subcontractor may demonstrate the SST scalable cluster in mid CY 1998 containing 0.5 Terabytes (TB) of memory and at least 25 Terabytes of RAID disk. The sum of peak and sustained performance on the sPPM demonstration code of the SST scalable cluster shall be at least four (4.0) TeraFLOP/s.

6.2.11 Scalable Development Environment Demonstration (MR)

The Subcontractor shall demonstrate the scalability of essential software capabilities in the application development environment across the clustered SMP system no later than the end CY 1998. Specifically, the Subcontractor shall use the sPPM demonstration code on the full cluster to demonstrate debugger, event tracing and performance statistics capabilities.

6.2.12 Sustained Stewardship TeraFLOP (SST) Demonstration (MR)

The Subcontractor shall demonstrate the SST scalable cluster no later than the end CY 1998 containing 0.5 Terabytes (TB) of memory and at least 25 Terabytes of RAID disk. The sum of peak and sustained performance on the sPPM demonstration code of the SST scalable cluster shall be at least four (4.0) TeraFLOP/s.

6.2.13 SST Installation (MR)

The Subcontractor shall install and support this system at either LANL or LLNL, as directed, for at least two years following a successful demonstration.

6.2.14 FY99 Plan and Review (MR)

The Subcontractor shall provide a detailed plan of activities and deliverables for fiscal year 1999 for University review and approval in the first quarter of FY99.

6.2.15 Memory Installation (MO)

The Subcontractor shall, at the University's option, install SST system memory up to an additional 1.0 TB by the first quarter of CY 1999. If the University chooses to exercise this option, installation of the additional memory before or after SST installation shall be at the Subcontractor's discretion.

6.2.16 Disk Delivery (MR)

The Subcontractor shall deliver the remaining 50 Terabytes of disk within 12 months of SST delivery.

6.2.17 Faster Disk Delivery (TR)

The Subcontractor may deliver the remaining 50 Terabytes of disk within 6 months of SST delivery.

6.3 Performance Reviews (MR)

Quarterly performance reviews shall be conducted between the Subcontractor's corporate executives and the University. The Subcontractor shall submit a Quarterly Project Status Report at least five working days before each quarterly review. The report shall provide the status of all work breakdown structure tasks and milestones in the critical path. It shall also contain narrative descriptions of anticipated and actual problems, solutions, and the impact on the project schedule. Numbered action items shall be taken, assigned, logged, and tracked by the Subcontractor. The minutes of all project reviews shall be recorded in detail by the Subcontractor and provided to the University for approval within 5 working days after the review.

6.4 SST TeraFLOP/s sPPM Demonstration (MR)

The sPPM demonstration code is of special interest to ASCI because it solves important hydrodynamics problems using an aggressive and demonstrated highly-efficient SMP cluster implementation of numerical methods relevant to DOE's Stockpile Stewardship Program. The sPPM demonstration code represents the current state of an ongoing effort which has demonstrated good processor performance, excellent multitasking efficiency, and excellent message passing parallel speedups all at the same time. Its use as an SST demonstration is to validate the ASCI effort by demonstrating a sustained TeraFLOP/s computation across all processors of a cluster of SMPs for a single large application which is of direct interest to the DOE Stockpile Stewardship Program as well as to scientific simulation in general. It is expected to bring recognition to the ASCI Program, to the Subcontractor and the University, as well as to the larger scientific high performance computing community.

The Subcontractor shall provide a best effort to achieve a sustained one (1.0) TeraFLOP/s execution rate of the sPPM benchmark code across the entire SST. The University shall witness, verify and certify the demonstration. A sustained performance rate shall be computed only for the time step update portion of the sPPM code for a problem size of approximately 2000-cubed grid points for a number of time steps sufficient to run for at least one (1.0) hour of wall clock time. The exact problem size will necessarily be determined by the actual SST cluster size, the available application memory, and the achieved sustained computation rate. The computation rate shall be determined based on the actual executed operations as shown below and the elapsed wall clock time required to complete the time steps on the SST system.

+, -, *	1 FLOP each
/, sqrt	4 FLOPs
min, max, abs, sign	1 FLOP each

The sPPM one (1.0) TeraFLOP/s SST demonstration code shall be derived from the sPPM code from the ID benchmark suite. It shall be programmed in Fortran with some C and shall contain no custom assembly language. It may use either single or double precision IEEE arithmetic (or even a mixture). It shall use POSIX threads and MPI message passing. It may use general or scientific library routines if they are, or will be, part of a supported library. It shall use the initial conditions built into the ID code without the image or texture maps. The time for initialization, visualization and restart dumps will not be included in the TeraFLOP/s rate determination - visualization and restart dumps could even be disabled. The timing prints per double time step can also be disabled but the one line “courant and energy” print is required on at least node 0 to either stdout or to the output file.

Optimizations will be allowed so long as they don’t specialize the functionality represented by the reference sPPM benchmark implementation. In particular, the source code for the hydro kernel of the sPPM benchmark code (i.e. the file sppm.m4) shall be used as provided by the University (i.e. unmodified), except for the addition of compiler directives. The University will continue its efforts to improve the efficiency of the code. Even for the subroutines in sppm.m4, tuning can be achieved through selecting alternate pieces of provided code through preprocessor flags, through the parameter IQ, etc. The goal is to emphasize higher level optimizations as well as compiler optimization technology improvements while maintaining “readable” code and physics modules. One obvious permitted modification in the parallel part of the sPPM demonstration code may be the overlapping of computation and communication (i.e., asynchronous communication).